

(19) World Intellectual Property Organization  
International Bureau



(43) International Publication Date  
17 October 2002 (17.10.2002)

PCT

(10) International Publication Number  
**WO 02/082214 A2**

(51) International Patent Classification<sup>7</sup>: **G06E**

(21) International Application Number: PCT/US02/10580

(22) International Filing Date: 5 April 2002 (05.04.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:  
60/282,028 6 April 2001 (06.04.2001) US

(71) Applicant: **PREDICTIVE NETWORKS, INC.**  
[US/US]; 689 Massachusetts Avenue, Suite 200, Cambridge, MA 02139 (US).

(72) Inventor: **CERRATO, Dean, E.**; 2 Hawthorne Place #2B, Boston, MA (US).

(74) Agents: **VALLABH, Rajesh et al.**; Hale and Dorr LLP, 60 State Street, Boston, MA 02109 (US).

(81) Designated States (*national*): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW.

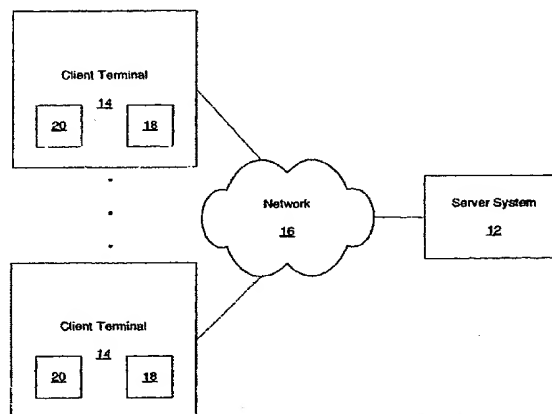
(84) Designated States (*regional*): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

**Published:**

— without international search report and to be republished upon receipt of that report

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: METHOD AND APPARATUS FOR IDENTIFYING UNIQUE CLIENT USERS FROM USER BEHAVIORAL DATA



(57) Abstract: A method and system are provided for identifying a current user of a terminal device from a group of possible users. The method includes providing a database containing multiple user input pattern profiles of prior user inputs to the terminal device. Each of the possible users of the group are associated with at least one of the user input pattern profiles. Current input patterns from use of the terminal device are detected. The current input patterns are combined and then dynamically matched with one of the user input pattern profiles, and the possible user associated with the matched user input pattern profile is selected as the current user. The system for identifying a current user of a terminal device from a group of possible users includes a database containing multiple user input pattern profiles of prior user inputs to the terminal device. Each of the possible users is associated with at least one of the user input pattern profiles. The system detects current input patterns from use of the terminal device, and then combines the patterns and dynamically matches the patterns with one of the user input pattern profiles. The system selects the possible user associated with the matched user input pattern profiles as the current user.

WO 02/082214 A2

## METHOD AND APPARATUS FOR IDENTIFYING UNIQUE CLIENT USERS FROM USER BEHAVIORAL DATA

### Related Application

The present application is based on and claims priority from U.S.

- 5      Provisional Patent Application Serial No. 60/282,028 filed on April 6, 2001 and entitled "Method and Apparatus for Identifying Unique Client Users from Clickstream, Keystroke and/or Mouse Behavioral Data."

### Background of the Invention

#### 10      Field of the Invention

The present invention relates generally to monitoring the activity of users of client terminal devices and, more particularly, to a method and system for identifying unique users from user behavioral data.

#### Description of Related Art

- 15      Various systems are available for profiling users of client terminal devices. Profiling typically involves determining demographic and interest information on users such as, e.g., age, gender, income, marital status, location, and interests. User profiles are commonly used in selecting targeted advertising and other content to be delivered to particular users. Delivery of targeted content is  
20      advantageous. For example, targeted advertising has been found to be generally more effective in achieving user response (such as in click-through rates) than advertising that is generally distributed to all users.

- Client terminal devices are commonly used by multiple individual users. For example, a home computer or a household television set can typically be  
25      expected to be used by various different family members at different times. Each particular user is likely to have a very different profile from other possible users, making delivery of targeted content ineffective. A need accordingly exists for distinguishing between various possible users at a given client terminal device.

### Brief Summary of Various Embodiments of the Invention

Certain embodiments of the present invention are directed to a method and system for identifying a current user of a terminal device from a group of possible users. The method in accordance with one embodiment includes providing a database containing multiple user input pattern profiles of prior user inputs to the terminal device. Each of the possible users of the group are associated with at least one of the user input pattern profiles. Current input patterns from use of the terminal device are detected. The current input patterns are combined and then dynamically matched with one of the user input pattern profiles, and the possible user associated with the matched user input pattern profile is selected as the current user.

The system for identifying a current user of a terminal device from a group of possible users in accordance with another embodiment includes a database containing multiple user input pattern profiles of prior user inputs to the terminal device. Each of the possible users is associated with at least one of the user input pattern profiles. The system detects current input patterns from use of the terminal device, and then combines the patterns and dynamically matches the patterns with one of the user input pattern profiles. The system selects the possible user associated with the matched user input pattern profiles as the current user.

These and other features of embodiments of the present invention will become readily apparent from the following detailed description wherein embodiments of the invention are shown and described by way of illustration of the best mode of the invention. As will be realized, the invention is capable of other and different embodiments and its several details may be capable of modifications in various respects, all without departing from the invention. Accordingly, the drawings and description are to be regarded as illustrative in nature and not in a restrictive or limiting sense with the scope of the application being indicated in the claims.

### Brief Description of the Drawings

For a fuller understanding of the nature and objects of various embodiments the present invention, reference should be made to the following detailed description taken in connection with the accompanying drawings wherein:

FIGURE 1 is a schematic diagram illustrating of a representative network in which a system in accordance with various embodiments of the inventions can be implemented;

FIGURE 2 is a flowchart generally illustrating the algorithm for matching a current clickstream with a stored clickstream profile in accordance with one embodiment; and

FIGURE 3 is a flowchart generally illustrating the fusion algorithm for identifying a unique user from multiple sources of input data in accordance with another embodiment of the invention.

### Detailed Description of Preferred Embodiments

The present invention is generally directed to a method and system for identifying a current user from a group of possible users of a client terminal device from user behavioral data. Once the user is identified, targeted content  
5 (such as, e.g., targeted advertising or recommended programming) can be delivered to the terminal device.

FIGURE 1 schematically illustrates a representative network in which a system for identifying unique users can be implemented. In general, the system  
10 includes a server system 12 for delivering content to a plurality of user terminals or client devices 14 over a network 16. Each user terminal 14 has an associated display device 18 for displaying the delivered content. Each terminal 14 also has a user input or interface interaction device 20 that enables the user to interact with a user interface on the terminal device 14. Input devices 20 can include, but are not limited to infrared remote controls, keyboards, and mice or  
15 other pointer devices.

In some embodiments, the network 16 can comprise a television broadcast network (such as, e.g., digital cable television, direct broadcast satellite, and terrestrial transmission networks), and the client terminal devices 14 can comprise, e.g., consumer television set-top boxes. The display device 18 can be,  
20 e.g., a television monitor.

In some embodiments, the network 16 can comprise a computer network such as, e.g., the Internet (particularly the World Wide Web), Intranets, or other networks. The server system 12 can comprise, e.g., a Web server, the terminal device 14 can comprise, e.g., a personal computer (or other Web client device),  
25 and the display device 18 can be, e.g., a computer monitor.

### Television System Embodiments

In the television system embodiments, the server system 12 can comprise, e.g., a video server, which sends data to and receives data from a terminal device 14 such as a television set-top box such as a digital set-top box.

5       The network 16 can comprise an interactive television network that provides two-way communications between the server 12 and various terminal devices 14 with individual addressability of the terminal devices 14.

10       The network 16 can, e.g., comprise a television distribution system such as a cable television network comprising, e.g., a nodal television distribution network of branched fiber-optic and/or coaxial cable lines. Other types of networked distribution systems are also possible including, e.g., direct broadcast satellite systems, off-air terrestrial wireless systems and others.

15       The terminal device 14 (e.g., set-top box) can be operated by a user with a user interface interaction device 20, e.g., a remote control device such as an infrared remote control having a keypad.

### Internet Embodiments

20       In the Internet (or other computer network) embodiments, the client terminals 14 connect to multiple servers 12 via the network 16, which is preferably the Internet, but can be an Intranet or other known connections. In the case of the Internet, the servers 12 are Web servers that are selectively accessible by the client devices. The Web servers 12 operate so-called "Web sites" and support files in the form of documents and pages. A network path to a Web site generated by the server is identified by a Uniform Resource Locator (URL).

25       One example of a client terminal device 14 is a personal computer such as, e.g., a Pentium-based desktop or notebook computer running a Windows operating system. A representative computer includes a computer processing

unit, memory, a keyboard, a pointing device such as a mouse, and a display unit. The screen of the display unit is used to present a graphical user interface (GUI) for the user. The GUI is supported by the operating system and allows the user to use a point and click method of input, e.g., by moving the mouse pointer on the display screen to an icon representing a data object at a particular location on the screen and pressing on the mouse buttons to perform a user command or selection. Also, one or more "windows" may be opened up on the screen independently or concurrently as desired. The content delivered by the system to users is displayed on the screen.

Client terminals 14 typically include browsers, which are known software tools used to access the servers 12 of the network. Representative browsers for personal computers include, among others, Netscape Navigator and Microsoft Internet Explorer. Client terminals 14 usually access the servers 12 through some Internet service provider (ISP) such as, e.g., America Online. Typically, multiple ISP "point-of-presence" (POP) systems are provided in the network, each of which includes an ISP POP server linked to a group of client devices 14 for providing access to the Internet. Each POP server is connected to a section of the ISP POP local area network (LAN) that contains the user-to-Internet traffic. The ISP POP server can capture URL page requests and other data from individual client devices 14 for use in identifying unique users as will be described below, and also to distribute targeted content to users.

As is well known, the World Wide Web is the Internet's multimedia information retrieval system. In particular, it is a collection of servers of the Internet that use the Hypertext Transfer Protocol (HTTP), which provides users access to files (which can be in different formats such as text, graphics, images, sound, video, etc.) using, e.g., a standard page description language known as Hypertext Markup Language (HTML). HTML provides basic document formatting and allows developers to specify links to other servers and files. These links include "hyperlinks," which are text phrases or graphic objects that

conceal the address of a site on the Web.

5 A user of a device machine having an HTML-compatible browser (e.g., Netscape Navigator) can retrieve a Web page (namely, an HTML formatted document) of a Web site by specifying a link via the URL (e.g.,  
www.yahoo.com/photography). Upon such specification, the client device makes a transmission control protocol/Internet protocol (TCP/IP) request to the server identified in the link and receives the Web page in return.

10 U.S. Patent Application Serial No. 09/558,755 filed April 21, 2000 and entitled "Method And System For Web User Profiling And Selective Content Delivery" is expressly incorporated by reference herein. That application discloses a method and system for profiling online users based on their observed surfing habits and for selectively delivering content, e.g., advertising, to the users based on their individual profiles.

#### User Identification from Behavioral Data

15 Various embodiments of the invention are directed to identifying a current individual user of a client device from a group of possible users. Such identification can be made from user behavioral data, particularly from input patterns detected in use of input devices (such as keyboards, mice, and remote control devices). As will be described in further detail below, the detected input  
20 patterns from a current user are compared with a set of input pattern profiles, which can be developed over time and stored in a database for the group of possible users. The current user is identified by substantially matching the current input pattern with one of the stored pattern profiles, each of which is associated with one of the possible users.

25 The database of input pattern profiles and software for detecting and matching current input patterns can reside at the client terminal 14 or elsewhere in the network such as at the server 12 or at the ISP POP server, or distributed at some combination of locations.

Different types of input data patterns can be used separately or in combination for identifying current users. The various types of input data patterns can include, e.g., (1) clickstream data; (2) keystroke data; (3) mouse usage data; and (4) remote control device usage data.

- 5 Briefly, in an Internet implementation, clickstream data generally relates to the particular websites accessed by the user. This information can include the URLs visited and the duration of each visit. In a television implementation, clickstream data generally relates to television surf stream data, which includes data on the particular television channels or programs selected by a user.
- 10 Keystroke data relates to keyboard usage behavior by a user. Mouse data relates to mouse (or other pointer device) usage behavior by a user. Remote control device data relates to usage behavior of a remote control device, particularly to operate a television set. For each of these types of user data, a sub-algorithm can be provided for detecting and tracking recurring patterns of user behavior.
- 15 User identification can be performed using any one of these types of user behavioral data or by combining two or more types of data. A so-called 'fusion' algorithm is accordingly provided for combining the outputs from two or more of the sub-algorithms to detect unique users. Briefly, the fusion algorithm keeps track of which combinations of particular patterns from, e.g., three types of data
- 20 (e.g., { click pattern "A," keystroke pattern "C," mouse pattern "F"}) recur most consistently, and associates these highly recurrent combinations with particular users.

#### Clickstream Behavior Tracking

- 25 Different Web or television users have different Web/television channel surfing styles and interests. The clickstream sub-algorithm described below extracts distinguishing features from raw clickstreams generated by users during Web/online surfing or television viewing sessions. In general, recurrent patterns of behavior in various observed different clickstreams are detected and

stored. Incoming clickstreams from a user client device are compared with these stored patterns, and the set of patterns most similar to the incoming clickstream pattern is output, along with their corresponding similarity scores.

### Clickstream Statistics

5           During an online session, the clickstream generated by the current user can be distilled into a set of statistics so that different clickstreams can be easily compared and their degree of similarity measured. This degree of similarity can be the basis for associating different clickstreams as possibly having been generated by the same person.

10           The following are sets of example clickstream statistics:

1) Total duration of visits to Top-N URLs or of viewing of Top-N television programs or channels

15           One set of clickstream statistics can be the top N (N can be variable, but is usually 8-10) unique URLs or channels/programs that appear in the current clickstream, selected according to total duration of visits at these URLs or of viewing of the channels/programs. The total duration is computed and stored. In addition to the Top-N unique URLs or channels/programs, a catch-all category named 'Other' can also be maintained.

2) Transition frequencies among Top-N URLs or channels/programs

20           Another set of clickstream statistics can be a matrix mapping 'From' URLs to 'To' URLs or 'From' channels/programs to 'To' channels/programs that captures the total number of all transitions from one URL or channel/program to the next in the clickstream. Transitions can be tracked among the Top-N URLs or channels/programs as well as those in the 'Other category.' In addition,  
25           'Start' and 'End' URLs or channels/programs can be used along with the 'From' and 'To' dimensions, respectively.

These statistics can be used to form a pattern of user surfing behavior. They can capture both the content of the clickstream (as represented by the content of the Top-N URLs or channels/programs), as well as some of the idiosyncratic surfing behavior of the user (as manifested, e.g., in transition behavior and proportion of sites or channels/programs that are 'Other').

User profiling can take into account the possibility that user input patterns are dependent on time. For example, in a television implementation, user viewing behavior can vary, e.g., based on the time of day or the given hour of a week.

## 10      Similarity Metrics

The similarity of clickstreams can be measured by calculating the similarity of statistics such as those described above between two different clickstreams. There are several different possible similarity metrics that can be used in distinguishing or comparing different clickstreams. Examples of such metrics include the following:

### 1) Dot-product of 'duration' unit vectors

Two 'duration' vectors are considered to be similar if they point in the same direction in URL or channels/programs -space, i.e., each clickstream visits many of the same Top-N URLs or channels/programs in similar proportions, regardless of the actual length or duration of the clickstream. This similarity is measured by computing the dot-product between the 'duration' unit vectors. Perfect similarity returns a value of unity, no similarity returns zero. The similarity of 'Other' values is preferably not included in this calculation since two clickstreams with identical 'Other' values might have in fact no similarity at all.

### 25      2) Dot-product of unit-vectorized 'transition' matrices

For similar reasons, the transition matrices can be compared using a dot-product metric. The matrices must first be vectorized (i.e., the elements are

placed in a vector). Transitions to and from 'Other' are considered generally significant and can be included in the calculation.

### 3) Similarity of 'Other' duration

5 The proportion of time spent at 'Other' URLs or channels/programs relative to the total user session time can be compared. Similarity is measured by computing for each of the two clickstreams to be compared the proportion of time spent at 'Other' URLs or channels/programs, then dividing the smaller of the two by the larger.

### 4) Similarity of total duration

10 This is a measure of similarity in the total duration of the clickstreams.

### 5) Similarity of total number of distinct URLs or channels/programs

This is a measure of the similarity in the total number of distinct URLs or channels/programs appearing in these clickstreams.

15 Each of these similarity metrics can be computed separately. If multiple metrics are used, they can be combined using various techniques. For example, a composite similarity metric can be computed by multiplying a select subset of these metrics together. Trial and error on actual data can determine which subset is most suitable. However, the similarity between duration vectors and transition matrices are likely to be more useful.

20 A good similarity metric will result in a good separation between users who have reasonably different surfing habits, and will not overly separate clickstreams generated by the same individual when that person is manifesting reasonably consistent surfing behavior.

### Matching Clickstreams Based on Similarity

25 Clickstreams that have high similarity values can be considered possibly

to have been generated by the same person. There are many possible ways to compute similarity. For example, one way to match a clickstream to one of a set of candidates is to select the candidate that has the highest similarity value. This technique is called a 'hard match'.

5           Alternatively, a somewhat more conservative approach can be to select a small group of very similar candidates rather than a single match. This group of candidates can subsequently be narrowed using some other criteria. This technique can be called finding a 'soft match'. A similarity threshold can be specified for soft matching. Soft matching is preferable when it is desired to  
10 match users according to multiple input pattern types such as keystroke and mouse dynamics in addition to clickstream behavior.

#### The Tracking Algorithm

It is desired to match incoming clickstreams with stored clickstream profiles representing recurrent clickstream patterns that have been observed  
15 over time. Each user is preferably associated with a single clickstream pattern profile. However, because an individual may have multifaceted interests, he or she may alternatively be associated with multiple clickstream pattern profiles. The process of matching incoming clickstreams with existing pattern profiles can be as follows as generally illustrated in FIGURE 2.

20           1) First at step 50, a set of recurrent clickstream profiles is created and stored in a database. It is expected that for a single client terminal there will be multiple different observed clickstreams generated by usually a small set of individual users, each of whom may have several different strong areas of interest that manifest in their surfing behavior. These can be represented as a set  
25 of clickstream pattern profiles that summarize the content and surfing behavior of most or all the observed clickstreams.

A clustering algorithm, e.g., can be used to generate a small set of clickstream pattern profiles to cover the space of observed clickstreams. New

clickstream profiles can be added whenever a new (i.e., dissimilar to existing profiles) clickstream is observed. Old profiles can be deleted if no similar incoming clickstreams have been observed for a given period of time. The growth/pruning behavior of the algorithm can be moderated by a similarity threshold value that determines how precisely the profiles are desired to match incoming clickstreams, and thus how many profiles will tend to be generated.

2) The next step 52 in the matching process is to dynamically (i.e., on-the-fly) match an incoming (i.e., current) clickstream to existing clickstream profiles. As a clickstream is being generated by a user, the partial clickstream can be compared on-the-fly at generally any time with the existing set of stored clickstream profiles. A hard or soft match can be made in order to determine the identity of the current user.

3) Next at step 54, the stored clickstream profiles are preferably retrained with data from completed clickstream. Upon termination of the current clickstream, the set of clickstream profiles is preferably retrained to reflect the latest clickstream observation. Clickstream profiles can be adjusted according to their similarity to the current clickstream.

### Keystroke Behavior Tracking

Another type of distinguishing user input pattern relates to the typing styles or keystroke dynamics of different users. Different users have different typing styles. A keystroke dynamics algorithm is accordingly provided to capture these different styles so that they may be associated with unique users.

The keystroke algorithm can be similar to the clickstream algorithm described above. The process can generally include the following steps:

- 1) Statistics on current keyboard activity occurring concurrently with the current clickstream are compiled.
- 2) A set of keystroke profiles based on past observations of keyboard

activity for a given terminal device are created and stored in a database.

3) The current keyboard activity is compared to the set of keystroke profiles to predict the user identity.

4) The keystroke profiles are preferably updated with the current  
5 keyboard activity once it has terminated.

Keystroke statistics can comprise a vector of average behavior that can be tested for similarity to other such vectors. The keystroke profiles for users can be created and trained in a similar manner as clickstream profiles. In addition, on-the-fly matching (hard or soft) of keystroke profiles to current keyboard input  
10 can be done in a similar manner as for clickstream matching.

One type of keystroke statistic that is particularly efficient and useful for characterizing typing behavior is the "digraph" interval. This is the amount of time it takes a user to type a specific sequence of two keys. By tracking the average digraph interval for a small set of select digraphs, a profile of typing  
15 behavior can be constructed.

The following is a list of frequent digraphs used in the English language (with the numbers representing typical frequency of the digraphs per 200 letters):

	TH	50	AT	25	ST	20
20	ER	40	EN	25	IO	18
	ON	39	ES	25	LE	18
	AN	38	OF	25	IS	17
	RE	36	OR	25	OU	17
	HE	33	NT	24	AR	16

IN	31	EA	22	AS	16
ED	30	TI	22	DE	16
ND	30	TO	22	RT	16
HA	26	IT	20	VE	16

5        Several of the most frequent digraphs can be selected for use in each keystroke profile. It is preferable that the digraphs be selected such that substantially the entire keyboard is covered.

#### Mouse Dynamics Tracking

10        There is generally an abundance of mouse (or other pointing device) activity during a typical web browsing session, making it useful to characterize user behavior according to mouse dynamics alone or in combination with other user input behavior.

15        Similar to keystrokes, the statistics collected for mouse dynamics can form a vector that can be compared for similarity to other such vectors and can be input to an algorithm such as a clustering algorithm to identify recurring patterns.

      Mouse behavior can include pointer motion and clicking action. The following are some examples of possible mouse usage statistics that can be gathered.

20        For clicking action, the average double-click interval can be determined. This can be the average time it takes a user to complete a double-click, which can be as personal as a keystroke digraph interval.

25        Also, for user clicking action, the ratio of double- to single-clicks can be determined. Much web navigation requires only single-clicks, yet many web users have the habit of double-clicking very frequently, which can be a

distinguishing factor.

For user pointer motion behavior, the average mouse velocity and average mouse acceleration statistics can be distinctive characteristics of users. Motion is preferably gauged as close to the hand of the person as possible since mouse ball  
5 motion is generally a more useful statistic than pixel motion.

Furthermore, the ratio of mouse to keystroke activity can also be a useful distinguishing characteristic of users. Some people prefer to navigate with the mouse, while others prefer use of a keyboard.

The algorithm for matching current mouse dynamics statistics with stored  
10 mouse usage profiles can be similar to that described above with respect to the clickstream algorithm.

#### Other Input Device Usage Tracking

Various other user input behavior can be used for determining unique users. For example, in the television embodiments, user input patterns can be  
15 determined from usage of devices such as infrared remote control devices. The following are examples of various usage characterizing patterns for such devices. These include (1) the length of time a button on the remote control device is depressed to activate the button control; (2) the particular channels selected for viewing; (3) the digraphs for selecting multi-digit channels; (4) the frequency of  
20 use of particular control buttons such as the mute button; and (5) the frequency with which adjustments such as volume adjustments are made.

The algorithms for matching statistics such as these to stored input profiles can be similar to those previously described.

#### The Fusion Algorithm

25 Multiple independent sources of user information (clickstream, keystroke, mouse and any other input data) can be available, each having a corresponding

algorithm that tracks recurring patterns in the input data. The existence of a set of unique users can be inferred from significant associations among these recurring input patterns. For example, a certain individual will tend to generate a particular keystroke pattern, a particular mouse pattern, and one or possibly several clickstream patterns. By detecting that these patterns tend to occur together and making an association, the existence of a unique user can be inferred.

A 'fusion' algorithm, which is generally illustrated in FIGURE 3, is provided to track associations among recurring patterns, to determine which patterns are significant, and to assign unique user status to those that are most significant. In addition, the fusion algorithm manages the addition and deletion of unique users from the system.

As previously described, each individual algorithm (e.g., for clickstream, keystroke, and mouse usage data) can perform a soft match between the current input data and its set of tracked patterns, and returns a list of most similar patterns along with their respective similarity scores as indicated in step 80. For example, the clickstream algorithm might return the following set of matched pattern data for clickstream data: { (pattern "2," .9), (pattern "4," .7), (pattern "1," .65) }, where the first entry of each matched pattern data indicates the particular matched pattern, and the second entry indicates the similarity score for that match.

The fusion algorithm tracks the frequency of recurrence of each possible combination of patterns among multiple individual algorithms. For example, a possible combination can have the form: { click pattern "c," key pattern "k," mouse pattern "m" }. If there are a total of C click patterns, K keystroke patterns, and M mouse patterns being tracked, then the total number of tracked combinations is  $C \times K \times M$ , which can be on the order of a few hundred, assuming the number of keystroke and mouse patterns is about 5, and the number of clickstream patterns is about 10.

Given a soft match from each of the tracking algorithms, a complete set of associations can then be enumerated at step 82 and scored at step 84.

Enumeration creates all the possible combinations. For each combination, a score is computed, which can be the product of the similarities of the members of the combination. It is not required that the score be the product of all the similarities; the score can also be based on various other possible combinations.

Then at step 86, an on-the-fly unique user identification can be made. The individual matching algorithms generate on-the-fly soft matches to current input data, which is then be used by the fusion algorithm to perform a hard match to its existing set of unique users to identify the current user.

Once the current user is identified, it is possible to effectively deliver to the user targeted content such as, e.g., targeted advertising or program viewing recommendations.

The fusion algorithm can then update the frequencies of recurrence for the enumerated combinations. One possible way of doing this would be by adding the current score of each particular combination to its previous cumulative score at step 88. It is preferable to decay all existing scores prior to applying the updates, so that infrequent or inactive patterns are weighted less heavily.

Unique users can be associated with patterns whose scores stand out significantly from the rest. After every update of combination scores, the fusion algorithm can determine at step 90 if any additions or deletions from the current set of inferred unique users is indicated. A new user can be added if the score of some combination exceeds a given threshold. An existing user can be deleted if the score of the corresponding combination falls below a given threshold. These thresholds are relative to the magnitudes of the entire set of scores, and represent degrees of "standing out" among all the combinations.

Before an addition occurs (entailing the creation of a new profile), it is preferably determined whether or not the presumed new user in fact

corresponds to an existing user. Since there is the possibility that an individual user could manifest more than one type of clickstream behavior, a new user having the same keystroke and mouse behavior of an existing user can be associated with the existing user, since keystroke and mouse behaviors are more likely to correlate strongly with individual users compared to clickstream behavior.

If an addition and a deletion occur at about the same time, it is possible that a particular user has simply "drifted", in which case that profile should be reassigned rather than being deleted and a new personal profile created.

While the embodiments described above generally relate to identifying or tracking individual users of client terminals, they can be applied as well to identifying recurring groups of such users. For example, television viewing at various times is performed by groups of individuals such as, e.g., by a husband and wife or by a group of children. These combinations of individuals could manifest distinct behavior.

For cases in which the system is unable to identify a user (or group of users) with a sufficient degree of certainty, the user could be designated as 'unknown' and an average user profile for the terminal device could be assumed.

Having described preferred embodiments of the present invention, it should be apparent that modifications can be made without departing from the spirit and scope of the invention.

### Claims

1. A method of identifying a current user of a terminal device from a group of possible users, comprising:

providing a database containing a plurality of user input pattern profiles of prior user inputs to said terminal device, each of said possible users being associated with at least one of said user input pattern profiles;

detecting at least one current input pattern from use of said terminal device; and

dynamically matching said at least one current input pattern with one of said user input pattern profiles, and selecting the possible user associated with the one of said user input pattern profiles as the current user.

2. The method of Claim 1 wherein said at least one current input pattern comprises a plurality of different input patterns, and wherein dynamically matching said at least one current input pattern comprises combining said plurality of different patterns and matching a combination of said different input patterns with one of said user input pattern profiles.

3. The method of Claim 1 further comprising retraining said plurality of user input pattern profiles in said database with said at least one current input pattern.

4. The method of Claim 1 further comprising determining a personal user profile associated with the current user.

5. The method of Claim 4 further comprising transmitting targeted content to said current user in accordance with said personal user profile.

6. The method of Claim 1 wherein said current input pattern comprises user clickstream data.

7. The method of Claim 6 wherein said clickstream data relates to

particular Web sites visited by the user or the duration of visits to the Web sites.

8. The method of Claim 1 wherein said current input pattern comprises user keystroke data.

9. The method of Claim 8 wherein said keystroke data comprises digraph interval data.

10. The method of Claim 1 wherein said current input pattern comprises user mouse usage data.

11. The method of Claim 1 wherein said current input pattern comprises user remote control usage data.

12. The method of Claim 1 wherein said terminal device comprises a computer.

13. The method of Claim 1 wherein said terminal device comprises a television set top box.

14. The method of Claim 1 wherein said steps are implemented in a computer, and said computer communicates with said terminal device over a network.

15. The method of Claim 14 wherein said network comprises the Internet.

16. The method of Claim 14 wherein said network comprises a nodal television distribution network.

17. A system for identifying a current user of a terminal device from a group of possible users, comprising:

a database containing a plurality of user input pattern profiles of prior user inputs to said terminal device, each of said possible users being associated with at least one of said user input pattern profiles;

means for detecting at least one current input pattern from use of said terminal device; and

means for dynamically matching said at least one current input pattern with one of said user input pattern profiles, and selecting the possible user associated with the one of said user input pattern profiles as the current user.

18. The system of Claim 17 wherein said at least one current input pattern comprises a plurality of different input patterns, and wherein said means for dynamically matching said at least one current input pattern combines said plurality of different patterns and matches a combination of said different input patterns with one of said user input pattern profiles.

19. The system of Claim 17 further comprising means for retraining said plurality of user input pattern profiles in said database with said at least one current input pattern.

20. The system of Claim 17 further comprising means for determining a personal user profile associated with the current user.

21. The system of Claim 20 further comprising means for transmitting targeted content to said current user in accordance with said personal user profile.

22. The system of Claim 17 wherein said current input pattern comprises user clickstream data.

23. The system of Claim 22 wherein said clickstream data relates to particular Web sites visited by the user or the duration of visits to the Web sites.

24. The system of Claim 17 wherein said current input pattern comprises user keystroke data.

25. The system of Claim 24 wherein said keystroke data comprises digraph interval data.

26. The system of Claim 17 wherein said current input pattern comprises user mouse usage data.

27. The system of Claim 17 wherein said current input pattern comprises user remote control usage data.

28. The system of Claim 17 wherein said terminal device comprises a computer.

29. The system of Claim 17 wherein said terminal device comprises a television set top box.

30. The system of Claim 17 wherein said system is implemented in a computer, and said computer communicates with said terminal device over a network.

31. The system of Claim 30 wherein said network comprises the Internet.

32. The system of Claim 30 wherein said network comprises a nodal television distribution network.

33. A computer system for identifying a current user of a terminal device from a group of possible users, comprising:

memory for storing a program and a plurality of user input pattern profiles of prior user inputs to said terminal device, each of said possible users being associated with at least one of said user input pattern profiles; and

a processor operative with the program to:

(a) detect at least one current input pattern from use of said terminal device; and

(b) dynamically match said at least one current input pattern with one of said user input pattern profiles, and selecting the possible user associated with the one of said user input pattern profiles as the current user.

34. A method of delivering targeted content to a current user of a terminal device used by a plurality of possible users, comprising:

providing a database containing a plurality of user input pattern profiles of prior user inputs to said terminal device, each of said possible users being associated with at least one of said user input pattern profiles;

detecting at least one current input pattern from use of said terminal device;

dynamically matching said at least one current input pattern with one of said user input pattern profiles, and selecting the possible user associated with the one of said user input pattern profiles as the current user;

determining a personal user profile associated with the current user; and

transmitting targeted content to said current user in accordance with said personal user profile.

35. The method of Claim 34 wherein said targeted content comprises targeted advertising.

36. The method of Claim 34 wherein said targeted content comprises recommended program viewing choices.

37. The method of Claim 34 wherein said personal profile includes demographic or preference data on said current user.

38. The method of Claim 37 wherein said demographic or preference data includes data on at least one of user age, user sex, number of children, income, and geographic location.

39. The method of Claim 34 wherein said steps are implemented in a computer server.

40. The method of Claim 39 wherein said server comprises a video server.

41. The method of Claim 39 wherein said server comprises a Web server.

42. The method of Claim 34 wherein said terminal device comprises a set top box and a television monitor.

43. The method of Claim 34 wherein said terminal device comprises a personal computer.

44. A method of identifying a current user of a terminal device from a group of possible users, comprising:

detecting a plurality of different types of current input patterns from use of said terminal device by a current user;

performing a soft match of each of said plurality of different types of current input patterns with a plurality of stored input patterns for each of said types of input patterns, said stored patterns representing input patterns for the group of possible users of said terminal device, said soft matches generating scored possible matches for each of said different types of data;

determining possible combinations of said scored possible matches;

determining a score for each said combination; and

for the combination having the highest score, selecting a possible user associated with said combination as the current user.

45. The method of Claim 44 further comprising retraining said plurality of stored input patterns with said current input patterns.

46. The method of Claim 44 further comprising determining a personal user profile associated with the current user.

47. The method of Claim 46 further comprising transmitting targeted content to said current user in accordance with said personal user profile.

48. The method of Claim 44 wherein said different types of current input patterns include a user clickstream pattern.

49. The method of Claim 48 wherein said clickstream pattern relates to particular Web sites visited by the user or the duration of visits to the Web sites.

50. The method of Claim 44 wherein said different types of current input patterns include a user keystroke pattern.

51. The method of Claim 50 wherein said keystroke pattern includes digraph interval data.

52. The method of Claim 44 wherein said different types of current input patterns include user mouse usage data.

53. The method of Claim 44 wherein said different types of current input patterns include user remote control usage data.

54. The method of Claim 44 wherein said terminal device comprises a computer.

55. The method of Claim 44 wherein said terminal device comprises a television set top box.

56. The method of Claim 44 wherein said steps are implemented in a computer, and said computer communicates with said terminal device over a network.

57. The method of Claim 56 wherein said network comprises the Internet.

58. The method of Claim 56 wherein said network comprises a nodal television distribution network.

59. A system for identifying a current user of a terminal device from a group of possible users, comprising:

means for detecting a plurality of different types of current input patterns from use of said terminal device by a current user;

means for performing a soft match of each of said plurality of different types of current input patterns with a plurality of stored input patterns for each of said types of input patterns, said stored patterns representing input patterns for the group of possible users of said terminal device, said soft matches generating scored possible matches for each of said different types of data;

means for determining possible combinations of said scored possible matches;

means for determining a score for each said combination; and

means for selecting a possible user associated with combination having the highest score as the current user.

60. The system of Claim 59 further comprising means for retraining said plurality of stored input patterns with said current input patterns.

61. The system of Claim 59 further comprising means for determining a personal user profile associated with the current user.

62. The system of Claim 59 further comprising means for transmitting targeted content to said current user in accordance with said personal user profile.

63. The system of Claim 59 wherein said different types of current input patterns include a user clickstream pattern.

64. The system of Claim 63 wherein said clickstream pattern relates to particular Web sites visited by the user or the duration of visits to the Web sites.

65. The system of Claim 59 wherein said different types of current input patterns include a user keystroke pattern.

66. The system of Claim 65 wherein said keystroke pattern includes digraph interval data.

67. The system of Claim 59 wherein said different types of current input patterns include user mouse usage data.

68. The system of Claim 59 wherein said different types of current input patterns include user remote control usage data.

69. The system of Claim 59 wherein said terminal device comprises a computer.

70. The system of Claim 59 wherein said terminal device comprises a television set top box.

71. The system of Claim 59 wherein said system is implemented in a computer, and said computer communicates with said terminal device over a network.

72. The system of Claim 71 wherein said network comprises the Internet.

73. The system of Claim 71 wherein said network comprises a nodal television distribution network.

74. The method of Claim 6 wherein said clickstream pattern relates to particular programs or channels selected by the user or the duration of viewing of said programs or channels.

75. The system of Claim 22 wherein said clickstream data relates to particular programs or channels selected by the user or the duration of viewing of said programs or channels.

76. The method of Claim 48 wherein said clickstream pattern relates to particular programs or channels selected by the user or the duration of viewing of said programs or channels.

77. The system of Claim 63 wherein said clickstream pattern relates to particular programs or channels selected by the user or the duration of viewing of said programs or channels.

78. A method of identifying a current subset of users of a terminal device from a set of possible users, comprising:

providing a database containing a plurality of user input pattern profiles of prior user inputs to said terminal device, various subsets of said possible users being associated with at least one of said user input pattern profiles;

detecting at least one current input pattern from use of said terminal device by a current subset of users; and

dynamically matching said at least one current input pattern with one of said user input pattern profiles, and selecting the subset of users associated with the one of said user input pattern profiles as the current subset of users.

79. The method of Claim 78 wherein said at least one current input pattern comprises a plurality of different input patterns, and wherein dynamically matching said at least one current input pattern comprises combining said plurality of different patterns and matching a combination of said different input patterns with one of said user input pattern profiles.

80. The method of Claim 78 further comprising retraining said plurality of user input pattern profiles in said database with said at least one current input pattern.

81. The method of Claim 78 further comprising determining a personal user profile associated with the current subset of users.

82. The method of Claim 81 further comprising transmitting targeted content to said terminal device in accordance with said personal user profile.

83. The method of Claim 78 wherein said current input pattern comprises user clickstream data.

84. The method of Claim 83 wherein said clickstream data relates to particular Web sites visited by the current subset of users or the duration of visits to the Web sites.

85. The method of Claim 83 wherein said clickstream data relates to particular programs or channels selected by the subset of users or the duration of viewing of said programs or channels.

86. The method of Claim 78 wherein said current input pattern comprises user keystroke data.

87. The method of Claim 86 wherein said keystroke data comprises digraph interval data.

88. The method of Claim 78 wherein said current input pattern comprises user mouse usage data.

89. The method of Claim 78 wherein said current input pattern comprises user remote control usage data.

90. The method of Claim 78 wherein said terminal device comprises a computer.

91. The method of Claim 78 wherein said terminal device comprises a television set top box.

92. The method of Claim 78 wherein said steps are implemented in a computer, and said computer communicates with said terminal device over a network.

93. The method of Claim 92 wherein said network comprises the Internet.

94. The method of Claim 92 wherein said network comprises a nodal television distribution network.

95. A method of identifying a current subset of users of a terminal

device from a set of possible users, comprising:

detecting a plurality of different types of current input patterns from use of said terminal device by the current subset of users;

performing a soft match of each of said plurality of different types of current input patterns with a plurality of stored input patterns for each of said types of input patterns, said stored patterns representing input patterns for various subsets of possible users of said terminal device, said soft matches generating scored possible matches for each of said different types of data;

determining possible combinations of said scored possible matches;

determining a score for each said combination; and

for the combination having score indicating a substantial match, selecting a subset of users associated with said combination as the current subset of users.

96. The method of Claim 95 further comprising retraining said plurality of stored input patterns with said current input patterns.

97. The method of Claim 95 further comprising determining a personal profile associated with the current subset of users.

98. The method of Claim 97 further comprising transmitting targeted content to said terminal device in accordance with said personal profile.

99. The method of Claim 95 wherein said different types of current input patterns include a user clickstream pattern.

100. The method of Claim 99 wherein said clickstream pattern relates to particular Web sites visited by the current subset of users or the duration of visits to the Web sites.

101. The method of Claim 99 wherein said clickstream pattern relates to particular programs or channels selected by the subset of users or the duration

of viewing of said programs or channels.

102. The method of Claim 95 wherein said different types of current input patterns include a user keystroke pattern.

103. The method of Claim 102 wherein said keystroke pattern includes digraph interval data.

104. The method of Claim 95 wherein said different types of current input patterns include user mouse usage data.

105. The method of Claim 95 wherein said different types of current input patterns include user remote control usage data.

106. The method of Claim 95 wherein said terminal device comprises a computer.

107. The method of Claim 95 wherein said terminal device comprises a television set top box.

108. The method of Claim 95 wherein said steps are implemented in a computer, and said computer communicates with said terminal device over a network.

109. The method of Claim 108 wherein said network comprises the Internet.

110. The method of Claim 108 wherein said network comprises a nodal television distribution network.

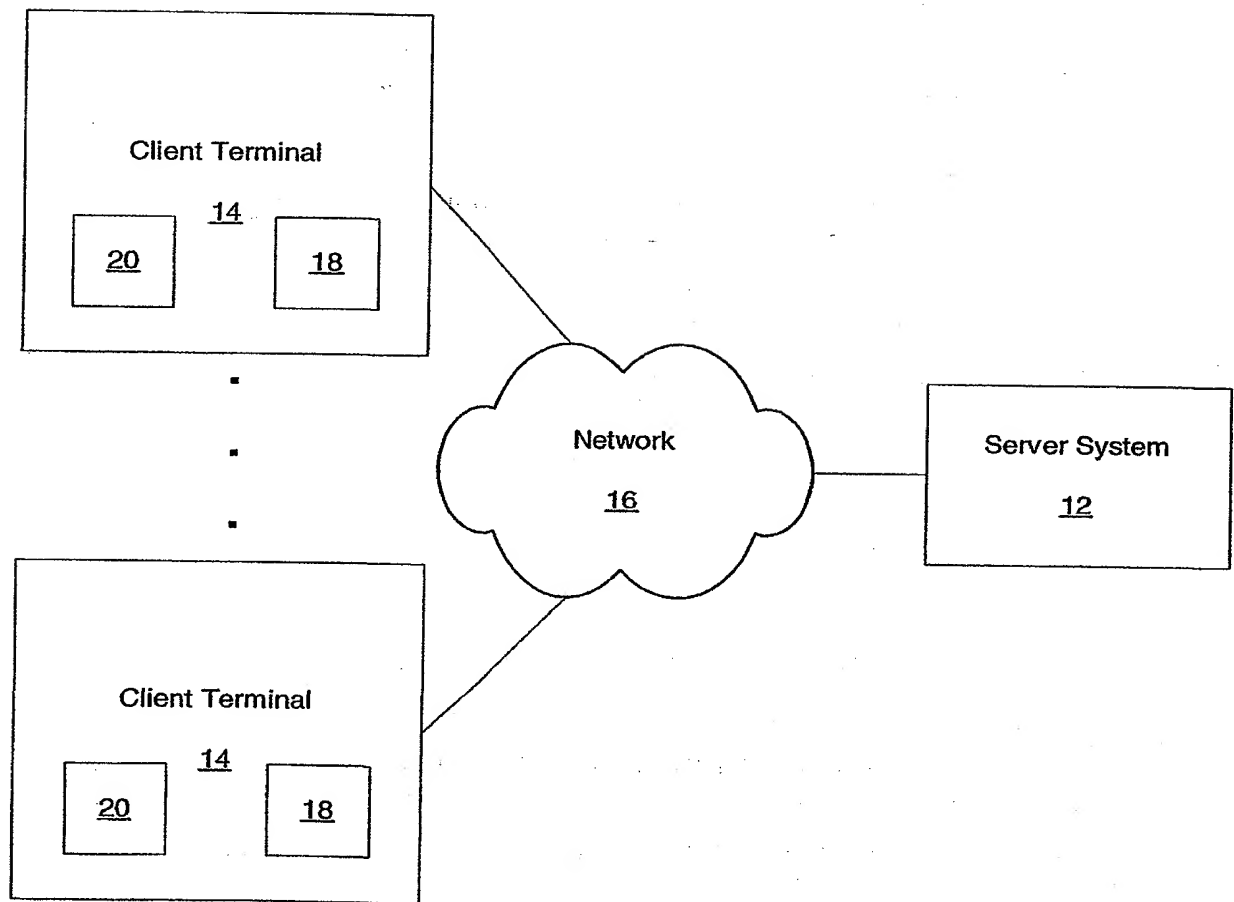


Fig. 1

2/3

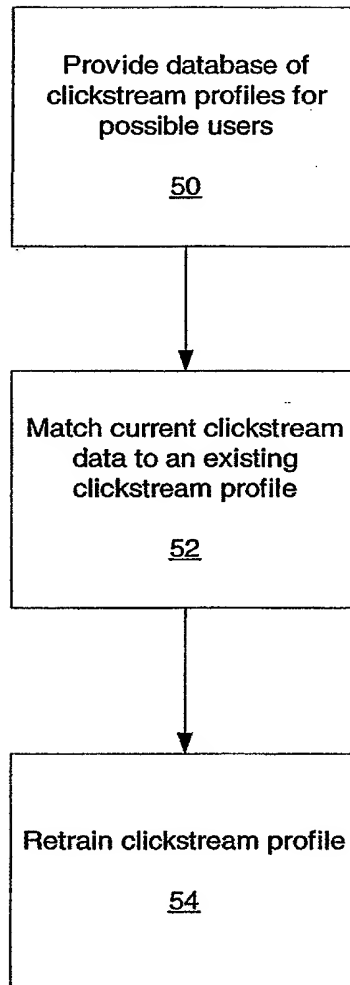


Fig. 2

3/3

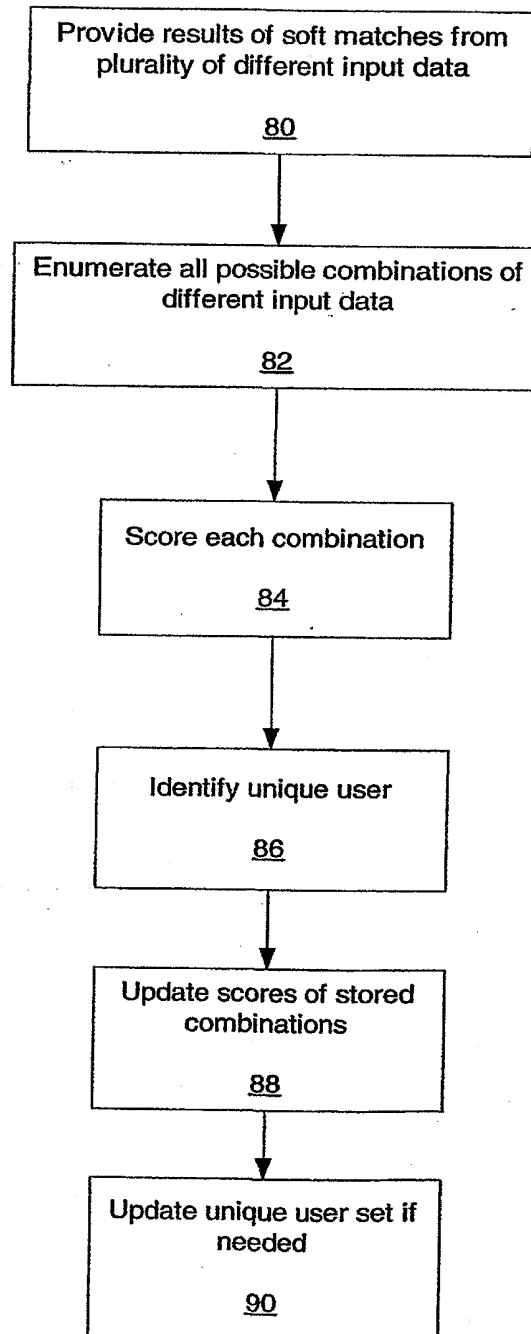


Fig. 3

(19) World Intellectual Property  
Organization  
International Bureau



(43) International Publication Date  
17 October 2002 (17.10.2002)

PCT

(10) International Publication Number  
**WO 2002/082214 A3**

(51) International Patent Classification<sup>7</sup>: **H04N 7/10**,  
7/14, 7/173, G06F 13/00

(21) International Application Number:  
PCT/US2002/010580

(22) International Filing Date: 5 April 2002 (05.04.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:  
60/282,028 6 April 2001 (06.04.2001) US

(71) Applicant: **PREDICTIVE MEDIA CORPORATION**  
[US/US]; 689 Massachusetts Avenue, Suite 200, Cam-  
bridge, MA 02139 (US).

(72) Inventor: **CERRATO, Dean, E.**; 2 Hawthorne Place #2B,  
Boston, MA (US).

(74) Agents: **VALLABH, Rajesh et al.**; Hale and Dorr LLP, 60  
State Street, Boston, MA 02109 (US).

(81) Designated States (*national*): AE, AG, AL, AM, AT, AU,  
AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU,

CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH,  
GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC,  
LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW,  
MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK,  
SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW.

(84) Designated States (*regional*): ARIPO patent (GH, GM,  
KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW),  
Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM),  
European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR,  
GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent  
(BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR,  
NE, SN, TD, TG).

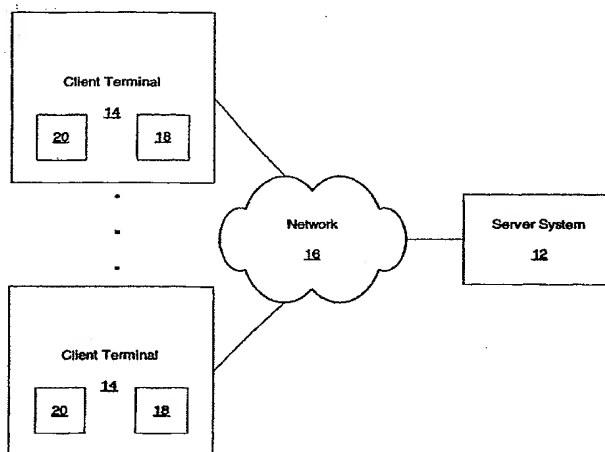
**Published:**

- with international search report
- before the expiration of the time limit for amending the  
claims and to be republished in the event of receipt of  
amendments

(88) Date of publication of the international search report:  
29 July 2004

*For two-letter codes and other abbreviations, refer to the "Guid-  
ance Notes on Codes and Abbreviations" appearing at the begin-  
ning of each regular issue of the PCT Gazette.*

(54) Title: METHOD AND APPARATUS FOR IDENTIFYING UNIQUE CLIENT USERS FROM USER BEHAVIORAL DATA



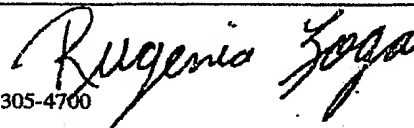
(57) Abstract: A method and system are provided for identifying a current user of a terminal device (14). The method includes providing a database containing multiple user input pattern profiles of prior user inputs to the terminal device (14). Each of the possible users of the group are associated with at least one user pattern profile. Current input patterns from the use of the terminal device (14) are detected, combined and then dynamically matched with one of the user input pattern profiles, and the possible user associated with the matched user input pattern is selected as the current user. The system for identifying a current user of a terminal (14) from a group of possible users includes a database containing multiple user input pattern profiles. The system detects current input patterns, then combines the patterns and dynamically matches the patterns with one of the user input pattern profiles. The system selects the possible user associated with the matched user input pattern profiles as the current user.

WO 2002/082214 A3

## INTERNATIONAL SEARCH REPORT

International application No.

PCT/US02/10580

<b>A. CLASSIFICATION OF SUBJECT MATTER</b> IPC(7) : H04N 7/10, 14, 173; G06F 13/00 US CL : 725/9, 34, 46 According to International Patent Classification (IPC) or to both national classification and IPC		
<b>B. FIELDS SEARCHED</b> Minimum documentation searched (classification system followed by classification symbols) U.S. : 725/9, 46  Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched  Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)		
<b>C. DOCUMENTS CONSIDERED TO BE RELEVANT</b>		
Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X — Y	US 5,758,257 A (HERZ, et al) 26 May 1998, Abstract; col. 26, lines 35-65; col. 25, lines 5-10; col. 37, lines 5-62; col. 41, lines 22-26; col. 51, lines 5-25	1-6, 13-21, 29-47, 55-62, 70-82, 91-98, 107-110  7-12, 22-28, 48-54, 63-69, 83-90 & 99-106
Y	WO 96/41494 A (COFFEY, et al) 19 December 1996, pg. 3, lines 15-24; pg. 12, lines 15-20; pg. 10, lines 25-30.	7-12, 22-28, 48-54, 63-69, 83-90 & 99-106
Y	US 5,796,952 A (DAVIS, et al) 18 August 1998, col. 8, lines 30-35; col. 13, lines 55-62.	7-12, 22-28, 48-54, 63-69, 83-90 & 99-106
Y	Other Measurement Methods; 07 October 1997, PC Magazine; 10/7/1997; pg. 2, 3rd paragraph.	7-12, 22-28, 48-54, 63-69, 83-90 & 99-106
<input type="checkbox"/> Further documents are listed in the continuation of Box C. <input type="checkbox"/> See patent family annex.		
<b>* Special categories of cited documents:</b>		
"A" document defining the general state of the art which is not considered to be of particular relevance	"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention	
"E" earlier application or patent published on or after the international filing date	"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone	
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art	
"O" document referring to an oral disclosure, use, exhibition or other means	"&" document member of the same patent family	
"P" document published prior to the international filing date but later than the priority date claimed		
Date of the actual completion of the international search 08 April 2004 (08.04.2004)	Date of mailing of the international search report 10 JUN 2004	
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US Commissioner for Patents P.O. Box 1450 Alexandria, Virginia 22313-1450 Facsimile No. (703)305-3230	Authorized officer Faile I. Andrew Telephone No. (703)305-4760 	

Form PCT/ISA/210 (second sheet) (July 1998)